speech and language processing 第十二章报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. p20的strong equivalence的定义如何理解？

讨论后的理解：equivalence被分为weak和strong两种，语言等价的意思是其产生的句子集合是相同的，而strong equivalence代表的是不仅仅句子集合相同，其内部的语法结构也相同，而weak equivalence代表的则是句子集合相同，但是生成结构有所差异。

1. 比起原始算法，CCG的优势是什么？

讨论后的理解：传统思路下的语法一般有很大的冗余，没有结构性，比较脆弱，规则也很多，并不适合计算机的使用。CCG的优势在于其更加具有逻辑性和结构性，其对应的规则简单并且是相互关联的。通过使用原子以及函数等概念，将句子的结构信息与单独的词语关联起来，从而不用生成复杂的语法规则，而是根据每个词语的位置及其范畴进行分析。

1. type raising的原理和意义是什么？

讨论后的理解：类型提升的价值在于可以将任意的简单范畴提升为函数的地位，这种情况下，其会让分析句子变得更加方便，比如与组合规则一起处理复杂的句子结构，比起只有基础规则的CCG更加广泛。其可以用在处理long-distance依赖问题上面。

1. 别人提出的问题的理解
2. P15页最后一段说的traces具体是什么意思吗？以及后面的例子能否解释一下具体是什么意思？

自己的理解：这个trace应该是说NONE标记，我认为它的用途可能是标注一些常见的长距离标注或者文中提到的语法运动，比如说he says...,这种情况是正常的，但是有的时候在英文使用习惯中，如果宾语太长的话，就会放在前面，把主语和谓语放在后面，加上none的话，一些解析器进行解析的时候看到none就会意识到句子的宾语在主谓前面，long distance问题也可以用相似的思路进行标记。

1. P17页12.4.2说的“flat”应该是什么意思？为什么会导致规则很长？

自己的理解：flat的意思应该是指浅显的，也就是直接根据语料总结出来的，相当于枚举出语法的规则。因为世界上的句子模式非常多，那么如果遇到某种结构非常奇特而且非常长的句子，其生成规则也要被总结，而且这种句子往往结构复杂，产生的规则必然也非常长。同时因为没有进一步归纳，这些规则之间会有冗余。

1. 读书计划

1、本周完成的内容章节：阅读完第12章。

2、下周计划：学习机器学习集成方法相关内容。

四、读书摘要总结：

1. Grammar Equivalence and Normal Form：

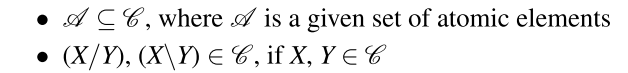
语法等价和范式，语言的一般定义是一个set of strings of words。而两个语言存在strong和weak equivalence两种情况，表示其等价的程度，但是无论程度强弱，其代表的集合应该是相同的。

范式则是语法产生式的一种规范形式，这在很多情况下有意义，比如编译原理。常用的范式有Chomsky normal form，其是binary branching，这样的好处是可以避免一些很长的语法规则，起到总结和生成的作用。

2. Lexicalized Grammars

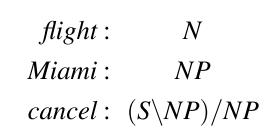
phrase-structure虽然简单直观，但是其也存在着很多问题，比如long-distance依赖问题，以及冗余的规则问题。为了克服这样的问题，提出了Combinatory Categorial Grammar，这种语法比起phrase-structure更加看重lexicon的作用，以此描述该语言。其基础是范畴，以及一系列归于范畴的词汇，和范畴在上下文中如何合并的规则，这三要素构成了CCG，并且使其有着广泛的应用。

范畴可以理解为一个原子元素或者一个单值函数：



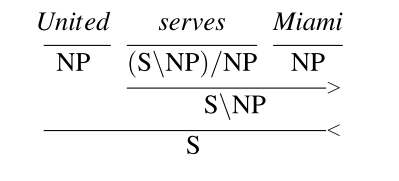
函数其实本质上应该可以理解为是一种对于上下文内容的理解机制，比如如果Y出现在了右面，那么这个词语就可以是X,这样的规则，而这种状态机结构非常适合计算机进行处理。

Lexicon可以理解为将范畴赋值给word得到的情况，比如说：



Rule规定了函数和其对应的参数是如何结合的。 

其可以在下面这样的例子中进行使用：

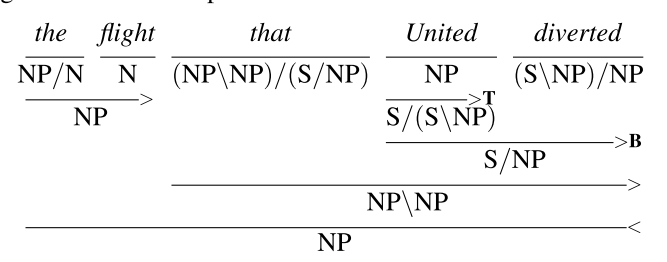


而为了进一步处理更加复杂的情况，引入了组合规则以及类型提升：





在这种情况下，其可以用来处理long-distance问题。



总结：本章介绍了基于模块的语法，主要介绍了常用于编译原理中的上下文无关文法以及更加看重单词作用的CCG。同时，其介绍了文法的作用，以及语法树和成分之间的依赖关系。虽然这些模型比较经典并且存在诸多问题，但是其非常具有逻辑性，并且从不同角度体现了语言的结构特征，比如说CCG就是从词语和其前后关联的内容中提取出其表达的类型等信息的，并且根据一条一条规则进行判断，实际上也是比较有效率的处理方法，这种基于规则的模型在过去没有足够数据和计算力的情况下也产生了很好的效果，这也是未来相关研究应该关注和借鉴的。